Foreign Language Optical Character Recognition, Phase II

Arabic and Persian Training and Test Data Sets

Final Report

Contract N00014-94-C-0182

Submitted to:

Dr. Thomas M. McKenna

Office of Naval Research

and

Dr. Sonny Maynard

Defense Advanced Research Projects Agency

By:

Robert B. Davidson and Richard L. Hopley

Science Applications International Corporation

McLean, Virginia 22102-3779

19970520 052

DTIC QUALITY INSPECTED 4

DISTRIBUTION STATEMENT A
Approved for public release
Distribution Unlimited

This report was presented as a paper at the 1997 Symposium on Document Image Understanding Technology, organized by the University of Maryland Institute for Advanced Computer Studies for the US Department of Defense, on 2 May 1997.

Foreign Language Optical Character Recognition, Phase II

Arabic and Persian Training and Test Data Sets

Final Report

Contract N00014-94-C-0182

Submitted to:

Dr. Thomas M. McKenna

Office of Naval Research

and

Dr. Sonny Maynard

Defense Advanced Research Projects Agency

By:

Robert B. Davidson and Richard L. Hopley

Science Applications International Corporation

McLean, Virginia 22102-3779

This report was presented as a paper at the 1997 Symposium on Document Image Understanding Technology, organized by the University of Maryland Institute for Advanced Computer Studies for the US Department of Defense, on 2 May 1997.

Foreign Language Optical Character Recognition
Phase II
Arabic and Persian Training and Test Data Sets

Robert B. Davidson and Richard L. Hopley

Science Applications International Corporation
McLean, Virginia 22102-3779

1. Introduction

This report describes the creation of large data sets consisting of bit-mapped images of real-world printed secular Arabic-alphabet text (in Arabic and Persian), accompanied by corresponding high-fidelity coded transcriptions (text ground truth), that have been systematically chosen, prepared, and documented. Each data set is divided into a training set, which is made available to developers, and a carefully matched equal-sized set of closely analogous samples, which is reserved for testing of the developers' products. The samples were systematically chosen to represent current vocabulary, usage, typography, and publication practices in major newspapers and news magazines, and in recent books and journals dealing with politics, economics, and commercial and military matters. Lexicons and character-frequency tables have been compiled for each data set and for the Arabic collection as a whole

Availability of these data sets is intended to catalyze development by others of Arabic and Persian optical character recognition (OCR) and natural language understanding systems. The tools developed to support this effort are readily applicable in other non-Roman alphabets. The expected end-users of such OCR systems are elements of the international business and foreign affairs communities that must communicate with and follow developments in parts of the world that communicate internally using non-Roman alphabets. At present, they are severely handicapped by their practical inability to use modern computer-based analytical tools. In particular, they find it difficult to build and use efficiently databases of current and archival information, because much of that information is available only in printed form in the local languages. Potential local compilers and users of collections of coded electronic information in these parts of the world operate under similar handicaps. Effective optical character recognition systems for these alphabets could remove or lower the most significant barrier to use of electronic information systems in these alphabets: the high cost (in resources and time) of converting printed information into coded electronic form. Because many of the needs of these communities are analytic (rather than contractual or legal), even imperfect OCR systems could contribute significantly.

Arabic is the official language of all of the countries of North Africa and most of the countries of the Middle East (including the economically important nations of the Arabian peninsula, and strategic Iraq, Libya, and Syria). Arabic is the sixth most

commonly used language in the world (after Mandarin Chinese, English, Hindi, Spanish, and Russian). In addition, the people of Iran, the Kurdish regions (in Iraq, Iran, Turkey, Syria, and the former Soviet Union), Afghanistan, and the Muslim portion of the Indian subcontinent (Pakistan and some parts of India) write their languages (Persian, Kurdish, Pashto, and Urdu,[1] respectively) in modified Arabic alphabets.

In the last few years, an increasing fraction of printed material in some of these countries (and in Iran) has begun to be produced electronically. However, only a small fraction of this electronically produced material is being published and disseminated in electronic form, and almost no archival material is available in electronic form. In many of these countries, electronic production of printed/typed material lags significantly, because the tools for electronic production in their alphabets (including capable desktop computers with native-language operating systems) are less well developed and less widely disseminated than the corresponding Roman-alphabet tools.

Efforts to date in Arabic alphabet OCR have identified and begun to address the problems involved in development of successful systems, but have not yet produced practical, widely usable systems. In part, this reflects inherent difficulties arising from particular characteristics of Arabic-alphabet languages (e. g., cursive connection of printed letters, and routine intrusion of ascenders and descenders into the "space" of adjoining letters; distinct characters that differ from one another only by the placement of small auxiliary elements; extensive shape changes of letters and combinations of letters depending on their environment within a word, etc.). But it also probably reflects limitations of the development efforts to date. In particular, no reported study has used a training data set of adequate size and diversity, or tested its recognition scheme against a large, real-world sample of printed Arabic (or Persian). Recognition rates of known systems for material outside their training data sets are unacceptably low. The availability of large, carefully chosen, widely accepted standard training and test data sets can support more rigorous (and ultimately more successful) development activities.

To support rigorous OCR system development efforts, our test and training data sets themselves meet severe standards of rigor. They are large enough (and carefully enough selected) to provide a population of words and characters that fairly represents the universe of material of most interest to a broad range of potential users. Real world samples (scanned with real scanners from real newspapers, magazines, and books) predominate. The correspondence between characters in the images and characters in the text truth files is very nearly perfect. All of these factors required diligent attention during development of the data sets, and careful configuration management and quality assurance. The rest of this report describes

---

[1] Urdu is essentially the same language (Hindustani) as Hindi, the official language of India. The major distinction is that Urdu is written in a modified Arabic alphabet and Hindi is written in the Devanagari alphabet.

our efforts to achieve these high standards in constructing two data sets: Arabic II and Persian I. (An Arabic I data set was developed in a previous effort.) These deliverables approach the standards described above much more closely than any comparable non-Roman alphabet data set that has been described in the literature.

2. Selecting the Sample Sources

For each language, highly qualified native scholars of the language first prepared analyses of printing and publication practices. They identified the most important font families, and patterns of use of those font families within different categories of publication and different "schools" of typography (e. g., the "Cairo" style of Arabic typography commonly followed in North Africa and the "Beirut" style of Arabic typography commonly followed in the Middle East). They documented standard usages within the character set of their language with respect to use of vowels, diacritical marks, numerals, special symbols, etc., which vary somewhat from country to country. Other experts (responsible for the African and Middle East collections of the Library of Congress) identified the most widely read and most influential regional and national newspapers, and also general interest and news magazines of wide circulation and some influence. Our language scholars made sure that the resulting selections broadly represented vocabulary, usage, typography, and publication practices. Among magazines/journals and books, they also selected titles to reflect bias in subject matter in favor of materials that dealt with current events, politics, and history; with military affairs; with science and technology; and with society and religion. (Literary works, low-level popular culture, and religious writings per se were purposely excluded.) Similar subject matter preference was applied later (see below) when choosing samples from the selected newspapers.

3. Sample Collection, Preparation, and Configuration Control

Most of our samples came from the extensive Arabic-alphabet holdings of the Library of Congress. The rest (mostly technical and other books, military affairs journals, and a few newspapers) came from material borrowed from private collections.

For both development and evaluation reasons, it is desirable to separate the pure Arabic/Persian character recognition problem from issues associated with page analysis and recognition in any language (the commercial OCR marketplace is addressing the latter issues aggressively). To support this separation, each of our samples is a single column of body text (with minor headings, but without headlines in display fonts, graphics, or tables) or a one-column book page. In order to avoid copyright infringement, we limited our sampling to at most four such insubstantial excerpts from any given issue of a newspaper, magazine, or journal, or from any given book. Native speakers made the final sample selections, consistent with the guidance of our language/typography experts, our source category targets, and our subject matter preferences. (The target distribution among source categories

was 50% newspapers, 30% magazines, and 20% books.) We took the excerpts in closely matched pairs, to help produce strictly comparable training and test data sets. (Within a training data set, it is also generally possible to select closely comparable pairs of samples, so that the developer can divide the data set into matching halves for training and in-house testing.) The average sample has about 300-350 words and about 1800-1900 characters. (See below for detailed statistics of the samples.)

Binary images were scanned at the collection sites, using a 600 dot per inch (dpi) flatbed scanner attached to a laptop computer. Flat settings were used on the brightness and contrast controls whenever the quality of the paper allowed (which was almost all the time for the sources selected). Scanned image originals were collected on floppy disks in Tag Image File Format (TIFF) (using lossless LZW compression); the original write-protected disks are retained as one of our primary archives. Serial numbers were assigned to the images as they were collected; a concurrent field log associated a full bibliographic citation (including English translation and Roman-character transliteration of the title) with the serial number. The field log entries were transcribed into our working catalog, which assigns each sample a unique identification number that incorporates much of the bibliographic information; the catalog allows each sample to be tracked through each of the subsequent steps of its processing. At the same time that the initial catalog entries were made, the images were transcribed from their floppy disk originals onto our working archival storage disks.

Each sample image was inspected. All extraneous graphic elements (e. g., unintentionally scanned remnants of headlines) impinging on the image area were cropped out or erased, as were rules, boxes, or areas of anomalously degraded text (e. g., smeared ink areas at the fold of a newspaper). One of us, who has had extensive professional experience in the printing industry, assigned each sample a print quality rating from 1 ("excellent") to 5 ("very poor"). Grade 5 samples, which were bad enough to give human readers difficulty, were excluded from further processing and from the final data sets. During the inspection, any unusual features of the samples, such as the presence of bullets, Roman or other non-standard characters, or bold or oblique type, were noted in their catalog entries.

4. Generation of Text Ground Truth

Printed versions of the images were assigned to two native-speaking typists. (Each typist received both same-size and twice-normal-size prints of each sample, the latter presumably being more readable.) Both contractual and practical measures were taken to ensure that the transcriptions would be truly independent. Each typist was carefully instructed to follow uniform conventions intended to produce identical typescripts from identical samples, with any typographical or other errors in the original images faithfully reproduced. Typists were instructed to substitute a place holder character for any non-Arabic-alphabet character that might occur in a sample.

After preliminary inspection (and, if necessary, correction) to bring the typescripts into closer conformance to our conventions, the two independent typescripts for each sample were reconciled by a native-speaking editor (assisted by custom text comparison software). Another copy of the printed images was used by the editor as the definitive version in these comparisons. At the conclusion of each editing session, the reconciliation software stored a corrected text truth file and recorded the errors made by each typist. (This same software is used in formal accuracy evaluations. See below.) A final visual inspection of the reconciled files brought the files into close visual conformance to conventions and detected (and corrected) some damage to the truth text files that occurred during the reconciliation process. Finally, a software-based inspection corrected violations of our conventions (Roman left-to-right space instead of Arabic right-to-left space, extra carriage returns, non-Arabic characters retained instead of replaced by the conventional place holder, etc.) that might be missed in a visual inspection.

5. Statistical Analysis and Assignment of Samples

The finished text truth files were parsed by software that counts the number of words and characters in each sample and in the data set as a whole. Tables 1 and 2 summarize this statistical information for our Arabic II and Persian data sets. In addition, for each alphabetical character, the number of times it occurred in the data set in each position (initial, medial, final, and free-standing) was also counted. When first encountered, each hitherto unseen word was recorded in a lexicon (as was the identity of the sample in which it first occurred[2]); when a word recurred, its count in the lexicon was incremented. (The software makes no attempt to group grammatical variants of the same word or root.[3]) The most recent version of the Arabic lexicon contains 69,905 words; the most recent version of the Persian lexicon contains 11,819 words.

The samples within each data set were divided into two closely matched subsets, a training subset and a test subset. Each subset has very nearly the same distribution of samples by source publication; *i. e.*, from a given book the same number of pages (one or two) are assigned to each subset; from a given newspaper or magazine/journal a very similar number of samples, yielding very nearly the same number of characters, are assigned to each subset. Care was also taken to ensure that each subset has a similar number of samples with each condition code. Within these constraints specific assignments were made in a way that balances the total number of characters in each set. Tables 3 and 4 summarize the results of this assignment process.

---

2 So that suspect entries can be examined readily in the original images.

3 Thus, the lexicons (along with their frequency-of-occurrence counts) are useful as they are for spelling checking, and are suitable input for more sophisticated grammatical analysis and word/root counting.

Table 1. Statistics for the Arabic II data set, by sample type.

	Distribution of Samples	Num. of Samples	Total Chars.	Chars/ Sample	Total Words	Words/ Sample
Newspaper	50%	259	465,033	1,795	76,288	294
Magazine	31%	159	296,985	1,867	48,240	303
Book	19%	99	153,797	1,553	24,808	103
Total Samples	100%	517	915,815	1,771	149,336	289

Table 2. Statistics for the Persian data set, by sample type.

	Distribution of Samples	Num. of Samples	Total Chars.	Chars/ Sample	Total Words	Words/ Sample
Newspaper	51%	348	647,111	1,859	121,794	349
Magazine	31%	212	374,693	1,767	71,321	336
Book	18%	126	220,399	1,749	41,913	332
Total Samples	100%	686	1,242,203	1,810	235,028	342

Table 3. Division of samples between training and test subsets for the Arabic II data set.

	Training			Test		
	Samples	Chars	Words	Samples	Chars	Words
Newspaper	132	232,527	40,928	127	232,506	41,612
Magazine	81	148,509	23,991	78	148,476	24,249
Book	50	76,868	12,425	49	76,929	12,383
Total Samples	263	457,904	74,406	254	457,911	74,939

Table 4. Division of samples between training and test subsets in the Persian data sets.

	Training			Test		
	Samples	Chars	Words	Samples	Chars	Words
Newspaper	174	323,604	121,655	174	323,507	122,460
Magazine	106	187,386	71,005	106	187,307	71,110
Book	63	110,187	41,870	63	110,212	41,956
Total Samples	343	621,177	117,265	343	621,026	117,763

These four tables summarize the data from two of the three the enclosed catalogs, the Persian Text Image (PTI) catalog and the Arabic Text Image II (ATI II) catalog. Also enclosed if the combined catalog of the two Arabic data sets created to date. The data from this third catalog is not included in this report because the work of the Arabic I data set was performed under a prior contract, and the results have already been reported. These catalogs each have five sections. The first section, the cover sheet, was our actual worksheet as the images were being collected, which was used to ensure that the proper distribution was followed. Then there are a number of pages of the individual catalog entries. Next, with each page headed "xTI File Stats" is a report generated by our Lexicon software, which parses each reconciled typescript, creates the Lexicon, and maintains statistics on each text file. Fourth is another report from the Lexicon software, in this case summarizing the results by document source (i.e., a given newspaper or magazine) and source type (newspaper, magazine, and book). Finally is a table giving the distribution of characters, by word position, throughout the entire data set.

6. Available Training Data Sets

We have made the training subsets of these data sets are available to developers, subject to approval by our sponsors. Generally, we request such approval whenever the developer is willing to describe the nature of the project in which the data set is to be used, and to agree not to redistribute the data set.

Along with the image and text truth files, the developer receives a catalog of the included samples. Each is identified by source type (newspaper, magazine/journal, book, or synthetic), source code (a unique identification number that allows the developer to know when samples came from issues of the same newspaper or magazine/journal), condition/print quality code (1=excellent, 2=good, 3=fair, 4=poor but readable), character count, and word count. The developer also receives a lexicon (which contains all words found in either the training or the test subsets and a word frequency table) and a character frequency file (which records how often each

character occurs, overall and in each position—initial, medial, or final—within words).

7. Formal Accuracy Testing

We have developed and (in one case so far) executed a protocol for formally testing OCR system accuracy. Using the testing data subsets and the reconciliation tools described above, we can help a developer (or a third party who is interested and has secured permission—through purchase of a commercial license or otherwise—to use a developer's system) to find out how well an Arabic or Persian recognition engine performs against a real-world data set, and what its specific strengths and weaknesses are.

First, we agree with the developer/user about what mutually convenient data transfer medium and protocols will be used for the test images and text files, and rehearse a data exchange. Then, at a mutually convenient time, we visit the developer's (or user's) facility to bring the test data set images and to witness the commencement of their processing. When the first ten images have been processed, the developer provides a copy of the resulting text files to support an initial on-site comparison of those files with the corresponding text truth files. Discussion of these initial results with the developer allows resolution of unforeseen issues at the early stages of (rather than after) the main effort of the evaluation.

When all of the test images have been processed, the developer delivers the rest of the resulting text files to us (in the same format and via the same medium as were used in the evaluation rehearsal). When practical (*i. e.*, when the processing can be completed within a day or two) we remain in the vicinity and return to the developer/user's site to pick up the files and observe the agreed clean-up procedure. In the latter, once we have received (and verified that we can read) all of the developer's text files, the developer/user removes all files associated with the testing from the developer/user's computing system. (The developer/user agrees not to re-create them from system backups.)

All of the developer's text files (or, if this is easier, all of our text truth files) are pre-processed to adjust for known differences in transcription conventions. Then the developer's text files are compared with our text truth files for the corresponding samples. Errors are counted by character, by type (insertion, omission, or replacement), and by position of the character within the word. As the results of these comparisons become available, we discuss them with the developer. Disagreements that the developer attributes to errors in our text truth files are investigated by humans, by reference to the original text images. For samples with reasonable numbers of errors (less than a few hundred[4]), our report details each of the developer's system's errors, summarizing them by type (character insertion,

---

[4] When the OCR system's output for a particular sample is found to bear little or no relation to the text in the image, reconciliation is discontinued.

deletion, or replacement), by character, and by character position. We calculate neither error rates nor other "scores," but provide information that allows the reader to readily calculate error rates or other metrics if desired.

Acknowledgments

We are grateful to Kamal J. Boullata and Dr. Shadin C. Shafa, who provided invaluable language and typography guidance. Dr. Paritchehr Coates was an extraordinarily conscientious and skillful editor. We are also grateful to the Library of Congress and some of its outstanding African and Middle East collection staff members (Beverly Gray, Dr. George Atieh, and Ibrahim Pourhadi), for allowing and helping us to use their collections.